Technical Appendix to accompany "Real-Time Evaluation of Email Campaign Performance"

André Bonfrer, Xavier Drèze

1. Selecting the hazard function

Researchers in marketing have used various specifications for parametric hazard function when doing survival analysis (see Jain and Vilcassim 1991, Sawhney and Eliashberg 1996; Chintagunta and Haldar 1998; Drèze and Zufryden 1998). Following their work, we considered the following four specifications: Exponential, Weibull, Log-Normal, and Log-Logistic.

We estimated a campaign level hazard rate model for each distribution using the complete set of opens and clicks available for each campaign (i.e., this is a straight hazard model that is neither split nor censored). We report the fit statistics for all four specifications in Table TA1 (open model) and Table TA2 (click model). The analysis suggests that the Log-Logistic distribution fits the data best overall for both open and click. The Log-Normal is a close second, but has the drawback of not having a closed form expression for its survivor function. It is important to note that the Exponential distribution performs relatively poorly, emphasizing the need for a non-constant hazard rate that allows for a delay between reception and open of an email, or between open and click (i.e., allows for enough time for consumers to process the message). The relatively poor fit of the Weibull distribution (which allows for a ramping up period) further shows that one also needs to accommodate for a decrease in the hazard rate after enough time has passed. Making the right assumptions regarding the change in hazard rate over time is thus crucial. This is especially true since much of the data available during the test will come from the first few hours of the test, representing the increasing part of the Log-Logistic hazard function. Estimating this based on a Weibull or Exponential hazard function would clearly mis-specify the model.

Another approach to identifying the proper hazard function specification is to fit a Box-Cox model. The specification of the Box-Cox hazard function is such that it nests many of the traditional hazard rate specification (Jain and Vilcassim 1991). Thus, one can fit a Box-Cox model and, assuming the true data generating process is one of the models nested in the Box-Cox specification, one can identify the proper hazard function by looking at the estimate parameters.

We fit the four parameter version of the Box-Cox described in Jain and Vilcassim (1991) to our campaigns. The parameter estimates failed to identify any specific distribution as a possible candidate (the model rejected the Exponential, Weibull, Gompertz, and Erlang-2). Unfortunately, the Log-Logistic is not one of the distributions that are nested within the Box-Cox specification (Chintagunta and Prasad 1998). Hence, the Box-Cox approach cannot validate our choice of the Log-Logistic as the base hazard specification; it can only validate our rejection of the other function.

We could have based our model on the Box-Cox hazard rate rather then the Log-Logistic. There are 3 reasons why we did not. First, the Box-Cox is a proportional hazard model. Hence, one cannot directly compare the likelihood function to the Log-Logistic likelihood for model selection purposes. Second, the Box-Cox has no closed formed solution for its density and survivor functions. This means that these two functions must be numerically integrated during the estimation. This considerably slows down estimation. The base hazard model (without censoring or split hazard) took up to 7 hours to estimate as opposed to less than a minute for the Log-Logistic. This defeats the purpose of developing a fast testing methodology. Third, the specification developed by Jain and Vilcassim (1991) is a four parameter function. We prefer our more parsimonious two parameter model as it puts less burden on the data.

Table TA1: Fit statistics for the virtual open time model. For the different specifications, the columns report the Log-Likelihood value (LL), the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Shaded cells show the lowest AIC and BIC for a specific campaign.

Campaign	Exponential			Weibull			Log-Normal			Log-Logistic		
	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC
1	-3,150	6,303	6,302	-2,702	5,412	5,408	-2,684	5,375	5,372	-2,722	5,451	5,448
2	-5,016	10,035	10,034	-4,142	8,292	8,288	-3,861	7,730	7,726	-3,868	7,744	7,740
3	-3,699	7,401	7,400	-3,100	6,208	6,204	-2,971	5,950	5,946	-2,951	5,910	5,906
4	-32,450	64,904	64,902	-26,424	52,858	52,852	-24,901	49,811	49,806	-25,015	50,039	50,034
5	-2,869	5,741	5,740	-2,457	4,922	4,918	-2,352	4,711	4,708	-2,361	4,729	4,726
6	-7,794	15,591	15,590	-6,418	12,844	12,840	-6,144	12,296	12,292	-6,163	12,334	12,330
7	-13,780	27,564	27,562	-12,145	24,299	24,294	-11,289	22,587	22,582	-11,268	22,545	22,540
8	-21,308	42,620	42,618	-18,126	36,261	36,256	-16,641	33,291	33,286	-16,613	33,235	33,230
9	-4,089	8,181	8,180	-3,328	6,664	6,660	-3,112	6,232	6,228	-3,115	6,238	6,234
10	-18,481	36,966	36,964	-15,523	31,055	31,050	-14,628	29,265	29,260	-14,540	29,089	29,084
11	-4,056	8,115	8,114	-3,576	7,160	7,156	-3,462	6,932	6,928	-3,483	6,974	6,970
12	-11,185	22,374	22,372	-9,793	19,595	19,590	-8,950	17,909	17,904	-8,918	17,845	17,840
13	-4,938	9,879	9,878	-3,998	8,004	8,000	-3,865	7,738	7,734	-3,863	7,734	7,730
14	-8,402	16,808	16,806	-7,126	14,260	14,256	-6,511	13,030	13,026	-6,473	12,954	12,950
15	-5,842	11,687	11,686	-5,223	10,454	10,450	-4,954	9,916	9,912	-4,897	9,802	9,798
16	-29,143	58,290	58,288	-25,437	50,884	50,878	-23,663	47,335	47,330	-23,651	47,311	47,306
17	-4,269	8,541	8,540	-3,701	7,410	7,406	-3,576	7,160	7,156	-3,596	7,200	7,196
18	-1,516	3,035	3,034	-1,244	2,495	2,492	-1,162	2,331	2,328	-1,155	2,317	2,314
19	-6,747	13,497	13,496	-5,651	11,310	11,306	-5,209	10,426	10,422	-5,216	10,440	10,436
20	-26,580	53,164	53,162	-23,524	47,057	47,052	-21,809	43,627	43,622	-21,847	43,703	43,698
21	-5,038	10,079	10,078	-4,061	8,130	8,126	-4,006	8,020	8,016	-3,990	7,988	7,984
22	-5,689	11,381	11,380	-4,997	10,002	9,998	-4,704	9,416	9,412	-4,688	9,384	9,380
23	-17,916	35,836	35,834	-16,204	32,417	32,412	-15,250	30,509	30,504	-15,317	30,643	30,638
24	-8,086	16,176	16,174	-6,735	13,478	13,474	-6,719	13,446	13,442	-6,698	13,404	13,400
25	-7,140	14,283	14,282	-5,711	11,430	11,426	-5,700	11,408	11,404	-5,706	11,420	11,416

Table TA2: Fit statistics for the virtual click time model. For the different specifications, the columns report the Log-Likelihood value (LL), the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Shaded cells show the lowest AIC and BIC for a specific campaign.

Campaign	Exponential			Weibull			Log-Normal			Log-Logistic		
	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC	LL	BIC	AIC
1	-117	236	236	-115	235	234	-109	223	222	-110	225	224
2	-217	436	436	-207	419	418	-189	383	382	-189	383	382
3	-216	434	434	-215	435	434	-191	387	386	-189	383	382
4	-2,567	5,137	5136	-2,506	5,019	5,016	-2,319	4,645	4642	-2,316	4,639	4,636
5	-144	290	290	-142	289	288	-126	257	256	-124	253	252
6	-711	1,425	1424	-697	1,400	1,398	-666	1,338	1336	-672	1,350	1,348
7	-938	1,879	1878	-908	1,823	1,820	-839	1,685	1682	-834	1,674	1,672
8	-1,972	3,947	3946	-1,921	3,849	3,846	-1,708	3,423	3420	-1,669	3,345	3,342
9	-337	676	676	-330	666	664	-310	626	624	-310	626	624
10	-546	1,095	1094	-546	1,098	1,096	-512	1,030	1028	-516	1,038	1,036
11	-844	1,691	1690	-842	1,690	1,688	-805	1,616	1614	-809	1,624	1,622
12	-396	794	794	-386	778	776	-347	700	698	-344	694	692
13	-517	1,036	1036	-506	1,018	1,016	-475	956	954	-475	956	954
14	-341	684	684	-340	686	684	-311	628	626	-307	620	618
15	-876	1,755	1754	-876	1,758	1,756	-836	1,678	1676	-838	1,682	1,680
16	-1,923	3,849	3848	-1,887	3,781	3,778	-1,713	3,433	3430	-1,693	3,393	3,390
17	-510	1,022	1022	-486	978	976	-441	888	886	-437	880	878
18	-101	204	204	-96	197	196	-89	183	182	-89	183	182
19	-232	466	466	-221	447	446	-202	409	408	-200	405	404
20	-851	1,705	1704	-840	1,686	1,684	-772	1,550	1548	-764	1,534	1,532
21	-122	246	246	-118	241	240	-110	225	224	-108	221	220
22	-465	932	932	-450	906	904	-412	830	828	-407	820	818
23	-938	1,879	1878	-930	1,867	1,864	-865	1,737	1734	-859	1,724	1,722
24	-556	1,114	1114	-537	1,080	1,078	-492	990	988	-486	978	976
25	-201	,404	404	-199	403	402	-180	365	364	-178	361	360

2. Accounting for different types of campaigns

Through the use of informative priors, our methodology builds on the information collected across all past campaigns. The more similar these campaigns are to the focal campaign, the better the methodology will perform. To the extent that additional (i.e., movie level) campaign data are available, it may be possible to improve the results by using different priors and different virtual time for different campaigns. Of course, when only a subset of campaigns is used to construct the prior, a trade-off must be made. On the one hand, more informative priors can be generated from using only campaigns that are similar to the focal campaign. On the other hand, there will be fewer campaigns to draw upon and thus the prior might be more diffuse.

To investigate the possibility of improving the performance of the model by building priors out of more similar campaigns, the email campaigns were first divided into three groups based on the genre of the movie they are promoting: Action Movies (13 campaigns), Romance Movies (eight campaigns), and Others (four campaigns - most of the "Others" category are special interest emails promoting more than one title each). We then reran the whole analysis for each three sub-groups of campaigns using the simplified model. That is, we re-ran the simulated test for each campaign using only the similar campaigns as prior, and using virtual time based on only the campaigns in the same sub-group. Figure TA1 shows that the speeds of time for each of the three subgroups are very similar but show some discrepancies. For instance, the 'Others' group exhibits slower virtual time until about 11 a.m., then speeds up and overtakes the other two groups until midnight. The 'Romance' group shows the opposite pattern.

The average parameter estimates for the different campaign genres are reported in Table TA3. These results reveal that the response parameters differ significantly across movie genres. The 'Romance' movie campaigns produce much lower open rates than 'Action' and 'Others' campaigns (15.93% vs. 20.38% and 20.98% respectively). Romance campaigns also have lower click rates on average (8.13%), leading to the lowest overall click-through rate (1.34%). In contrast, 'Other' campaigns have much higher click rates (15.91%), leading to the highest overall click-through rates (3.80%).

Table TA3 also shows the average shape and location parameters for the Log-Logistic hazard rate of each group. To help with the interpretation of these parameters, we show the resulting hazard rates in Figure TA2. This graph shows that the 'Others' category has a much

flatter hazard rate. It peaks about one hour and 45 minutes after the email has been sent and after two days is only reduced by half. The 'Action' genre, in contrast, peaks in about 40 minutes and loses about 80% in two days. At that point, the hazard rate is actually lower than for 'Others.'

This shows that the 'Others' category is much more appealing to consumers and has more staying power. This might be due to the fact that they promote more than one movie and thus may have a wider appeal. 'Romance' campaigns, in comparison, are much less appealing. They garner lower open rates, lower click rates and have a much steeper hazard function.

These differences across types of campaigns are quite large and suggest that using the priors based on similar campaigns rather than the general priors would lead to faster and more accurate tests. However, in our simulated tests the sub-group models do not perform as well on average as the pooled model. This result can be explained by the reduced ability for shrinkage to work given the smaller number of campaigns used to build the prior. In the 'Others' category for instance, there are only 4 email campaigns. This means that when testing a campaign, only three other campaigns are used as priors. This reveals itself to be too few campaigns to produce helpful priors. The sub-group tests suggest that it is better to use priors based on the full set of 24 other campaigns even if the majority of these campaigns are relative to somewhat different types of offer. It is likely that the reason for this is that, even though these other campaigns promote different types of movies, they still promote movies and are still aimed at the same basic type of consumers and thus contain relevant information.

The lack of positive results in this sub-group analysis does not invalidate the premise. We do find significant differences in behavior across the three types of campaigns. Unfortunately, there are not enough campaigns for each campaign genre to make the sub-group analysis work. We have no doubt that, as the firm collects more data on subsequent campaigns, the sub-group analysis would be worthwhile. It is easy to see how it would be optimal for the firm to start by using overall priors and then switch to more targeted priors as it builds a larger portfolio of past campaigns.

Table TA3: Average Parameter Estimates for the Three Different Types of Campaigns

Genre:	**N**	δ_O	α_O	λ_O	δ_C	**CTR**
Action	13	20.38%	1.06	0.0018	8.58%	1.88%
Romance	8	15.93%	1.15	0.0020	8.13%	1.34%
Other	4	20.98%	1.17	0.0012	15.91%	3.80%
All campaigns	25	19.1%	1.07	0.0019	0.0847	1.62%

Figure TA1: Speed of Time for Three Different Types of Email Campaigns

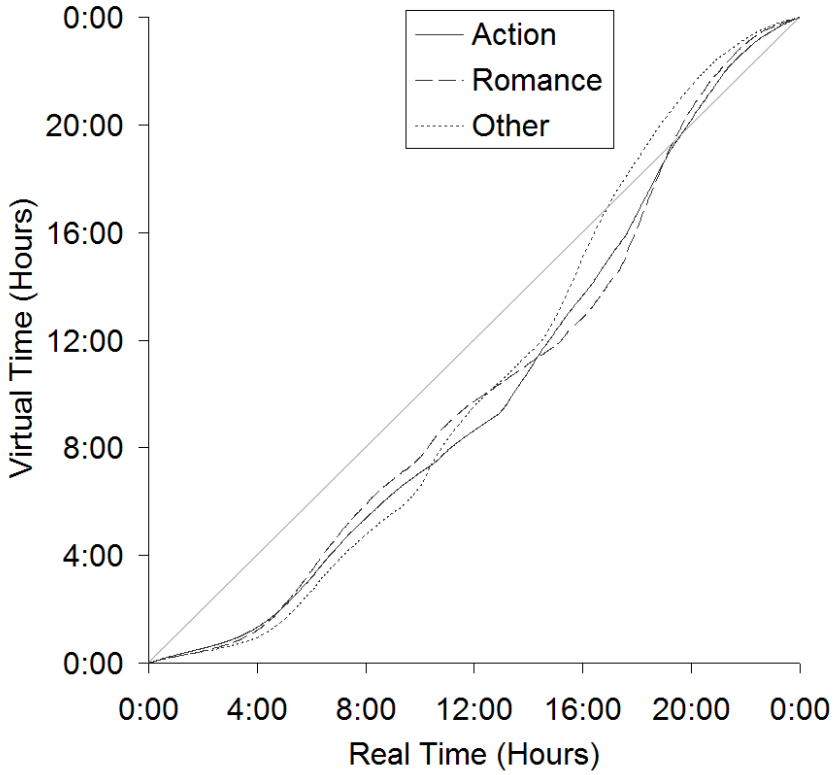
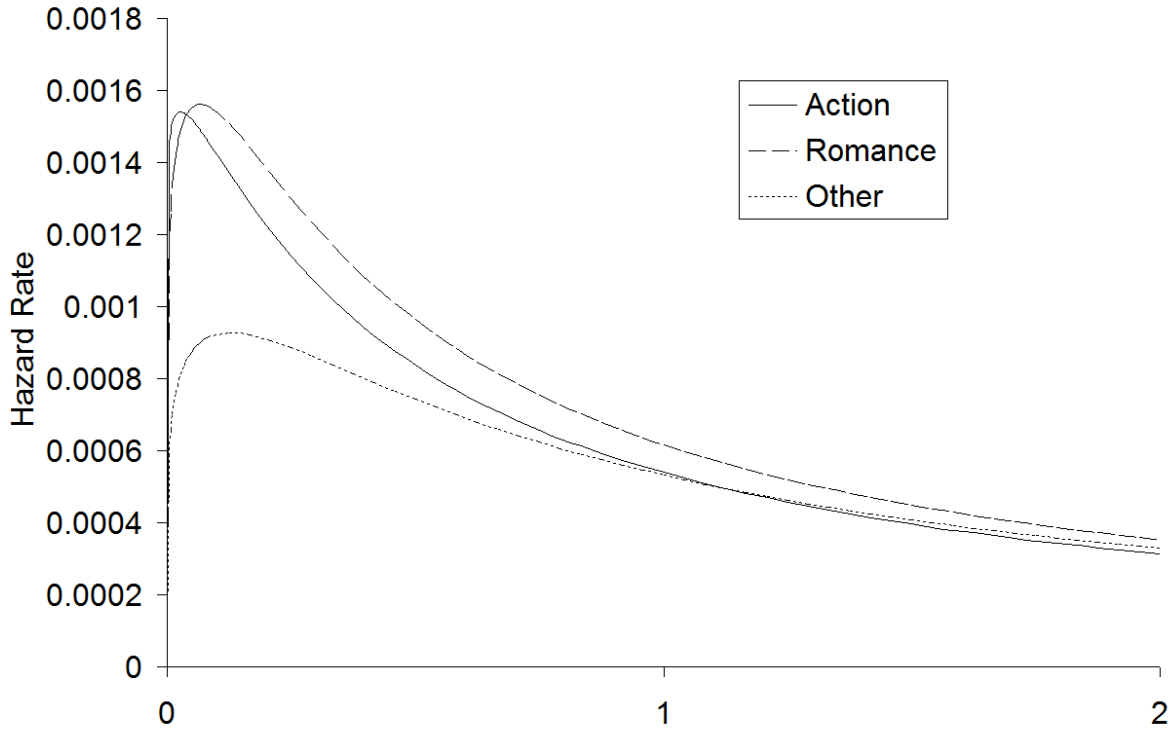


Figure TA2: Email campaign open hazard rates by genre



3. Rule 5: Asymmetrical Loss Function

In the paper, we propose to account for parameter uncertainty by adjusting the length of the test. Tests producing estimates with a large standard error would be run longer than tests that have small standard errors. Another way to incorporate the uncertainty around the point estimate into the manager's decision is to consider the fact that in many cases, under-estimation errors have a different impact on a firm's operations than over-estimation errors. In the case of email marketing, over- and under-estimation errors could be viewed from two perspectives. First, consider the low variable costs of sending email. In the short term, it is much more costly to not send a good campaign (loss opportunity) than to mistakenly send a bad one (cost of sending the campaign). Thus, under-estimation errors should be penalized more heavily than over-estimation errors and a manager should err on the side of sending too many campaigns rather than too few.

A second view is that a legitimate email firm operates on a permission basis. It cannot send emails to an individual who unsubscribe from the mailing list. Thus, a campaign manager might be concerned that under-performing emails may lead the recipients to re-evaluate the benefits they receive from the firm (maybe associating the emails with SPAM) and potentially defect. Such defection is costly as it deprives the firm from future revenues. In such case, the argument could be made that the relatively low cost of sending an email is more than matched by the high cost of defection and thus campaign managers should penalize over-estimation (leading to sending poor performing campaigns) more than under-estimation (leading to not sending well performing campaigns).

Given these two perspectives, the differential treatment of over- and under-estimation errors in the decision process can be handled using an asymmetrical loss function. Varian (1975) proposes the use of the LINEX loss function in such a case. This function is approximately linear on one side of zero (error) and approximately linear on the other side, depending on the value of a parameter a. The sign and magnitude of a represent the direction and degree of symmetry (when $a > 0$, overestimation is more serious than underestimation, and vice-versa.) When a is close to 0, the LINEX loss is almost symmetric and approximately equal to the squared error loss. Although other asymmetric loss functions exist, the LINEX loss function is one of the more commonly used in statistical applications (Calabria and Pulcini 1998; Pandey 1997; Zellner 1986).

If we call Δ the difference between the estimated and the true value of the parameter of interest then the LINEX loss function is expressed as:

$$l(\Delta) \propto \exp(a\Delta) - a\Delta - 1; a \neq 0.$$

As shown in Soliman (2002), the corresponding Bayesian estimator that minimizes the posterior loss is:

$$u^* = -\frac{1}{a}\log\left(E_u\left[\exp(-a.u)\right]\right).$$

In our context, *u* is the predicted CTR. To compute $E_u\left[\exp(-a.u)\right]$ we approximate the integral over *u* of $f(\tilde{u})\exp(-a.\tilde{u})$. To accomplish this, we make 10,000 draws from the posterior distribution of the estimated parameters (i.e., $\hat{\delta}_o, \hat{\delta}_c, \hat{\alpha}_o, \hat{\lambda}_o$) taking correlations among parameters into consideration. The CTR ($\tilde{u}$) resulting from such parameter draws can then be computed, which can then be used to compute the mean of $\exp(-a.\tilde{u})$. We adopt the perspective of a manager who is afraid of missing a good campaign so that a < 0; for exposition purposes, we used *a* = -10. Other values could be used depending on the risk aversion of the managers. The results are presented in Table A4.

Using the LINEX loss function (Rule 5), with the value of *a*=-10 (i.e., being afraid of missing a good campaign) leads to one more campaign being selected (as compared to the results of Rule 3). This is to be expected as the fear of missing a good campaign would lead to a more lenient test. The net result is a slight decrease in response rate as the campaign turns out to be an underperformer.

Table A4: Results from the Decision Rule Simulation

	Actual	**Rule 1** Doubling Method	**Rule 2** Horse Race	**Rule 3** Proposed model	**Rule 4** Adaptive model	**Rule 5** LINEX a=-10
Number of Campaigns Selected	7	11	14	7	6	10
True Positives		7	6	6	6	6
False Positive		4	8	1	0	4
False Negative		0	1	1	1	1
Average Testing Time (Hours)		14:00	2:30	2:30	2:57	2:30
Minimum Testing Time		14:00	2:30	2:30	1:00	2:30
Maximum Testing Time		14:00	2:30	2:30	6:00	2:30
Click-through Rate	4.45%	3.34%	2.66%	4.29%	4.77%	3.37%
Improvement over No Rule	123%	67%	33%	114%	139%	69%
Improvement over DM	33%	0%	-20%	28%	43%	1%
Revenue per Email sent	$.027	$.016	$.009	$.025	$.031	$.017
Revenue per Name (=Revenue per email sent × # of Campaigns)	$.190	$.172	$.122	$.178	$.183	$.168

4. Study of email newsletter frequency

To understand the extent to which companies are likely to benefit from real-time testing, we set out to get an indication of the number of companies who are engaged in frequently sending emails (e.g. once a week or more) to their permission-based lists. We subscribed a group of pseudo-recipients to 196 newsletters and monitored their activity for six months. These newsletters belonged to a wide range of companies (Airlines, Apparel, News, Electronics…). Since some companies may customize their newsletters we wanted to ensure there was some heterogeneity in the "recipients" we used. Thus whenever the registration process asked for gender information we registered different male and female accounts; when age information was asked, we registered both a 25 and a 45 year old. In addition, for each newsletter-age-gender combination, we registered three different recipients: one for which the incoming emails would be left untouched; one for which the emails would be opened; and one for which the emails would be opened and, if they contained links, those links would be clicked on to show interest in the content. Hence, for a web site that asks for both gender and age information, 12 recipient accounts would be opened (2x2x3). If the web site did not ask for any demographic information, only three accounts would be opened. Of the 196 newsletters, 16% asked for gender or age information (split evenly between the two), another 16% asked for both types of information. A total of 107 (55%) of the newsletters actually contacted our recipients with enough frequency to be considered as actively engaging in Permission-Based Marketing (we set the cut-off at three or more emails per account during the six-month period) for a total of 12,946 emails received. The level of customization among the active newsletters was low (see Table TA5). Four percent of the newsletters customized based on gender information (10% of the active newsletters that asked for gender information), five percent customized based on age (22% of the active newsletters that asked for age information), and five percent customized based on consumer actions (i.e., open or click). Finally, 61% of the newsletters used a fixed contact periodicity; weekly newsletters (63%) being the preferred contact interval (see Table TA6).

Table TA5: Description of Newsletters

	Subscribed		Received more than 3 emails from company		Received customized content	
N	196		107			
Gender information	48	24%	41	38%	4	10%
Age information	47	24%	23	21%	5	22%

Table TA6: Contact Frequency of Subscribed Newsletters

Periodicity	N	Percent
Day	1	2%
Bi-Weekly	3	5%
Weekly	41	63%
Bi-Monthly	14	22%
Monthly	6	9%
Total	65	

References for Technical Appendix

Calabria, R. and G. Pulcini (1996), "Point estimation under asymmetric loss functions for left-truncated exponential samples," *Communications in Statistics Theory and Methods*, 25 (3), 585–600.

Chintagunta, Pradeep K. and Sudeep Haldar (1998), "Investigating Purchase Timing Behavior in Two Related Product Categories," *Journal of Marketing Research*, XXXV (February), 43-53.

Chintagunta, Pradeep K. and Alok R. Prasad (1998), "An Empirical Investigation of the "Dynamic McFadden" Model of Purchase Timing and Brand Choice: Implications for Market Structure," Journal of Business & Economic Statistics, 16 (1), 2-12.

Drèze, Xavier and Fred Zufryden (1998), "A Web-Based Methodology for Product Design Evaluation and Optimization," *Journal of the Operation Research Society*, 49, 1034-43.

Jain, Dipak C. and Naufel J. Vilcassim (1991) "Investigating Household Purchase Timing Decisions: A Conditional Hazard Function Approach," *Marketing Science*, 10 (1), 1-23.

Pandey, B. N. (1997), "Testimator of the scale parameter of the exponential distribution using LINEX loss function," *Communications in Statistics Theory and Methods*, 26 (9), 2191–202.

Sawhney, Mohanbir S. and Jehoshua Eliashberg (1996), "A Parsimonious Model for Forecasting Box-Office Revenues of Motion Pictures," *Marketing Science*, 15 (2), 113-131.

Soliman, Ahmed A. (2002), "Reliability Estimation in a Generalized Life-Model with Application to the Burr-XII," *IEEE Transactions on Reliability*, 51 (3), 337-43.

Varian, Hal R. (1975), "A Bayesian Approach to Real Estate Assessment," in *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, eds. Stephen E. Fienberg and Arnold Zellner, Amsterdam: North-Holland, 195-208.

Zellner, Arnold (1986), "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," *Journal of the American Statistical Association*, 81 (394), 446-51.